A. FARAJ* I. C. LERMAN** N. VALETTE***

* DER/EDF/GRETS 30 rue de Condé 75006 PARIS
** I.R.I.S.A. Campus de Beaulieu 35042 RENNES
*** LAA/SLC/EVP C.N.E.T. Lannion 22301 LANNION

ORDINAL CODING AND CURVE CLASSIFICATION

1 - INTRODUCTION

The study presented in this paper was carried out on the basis of data measuring the disturbances caused by electrical phenomena. Mathematically, the data takes the form of curves defined over an interval T bounded by R with values in a part of R . They are temporal data. A certain number of methods for analysing this type of data exist, notable due to Deville who, under the name of harmonic analysis, proposed a generalisation of Principal Components Analysis (P.C.A.) to a family of random variables (Xt).

The curves can be considered either as the variables measured at the points t of T or as individuals on which these points t take measures We have here two assumptions that analyse the information contained in a corpus of curves in different ways. The first of them is a morphological analysis and the second is a chronological analysis of the deviations from the trend.

The only thing that our method has in common with harmonic analysis is the descriptive view that is specific to data analysis. It was born from the idea of classifying the curves in two stages.

- The first based on their shape, then within the same class of shape
- The second based on amplitude

The curves that initially considered as quantitative variables are next coded into ordinal qualitative variables (V.Q.O.). We then propose a similarity index between the variables from coding, after which we classify them using an " algorithme de la vraisemblance des liens " (A.V.L.) algorithm. This coding work consists in sampling the set of values reached by the curve at intervals corresponding to the modalities of the coding's qualitative variable.

Transforming a quantitative variable into a qualitative variable is obviously accompanied by a loss of information. This loss could be minimised by associating a modality with each quantity that the quantitative variable attains. The disadvantage would be that we would obtain qualitative variables having a very high number of modalities. The use of A.V.L. is frequent on V.Q.O.s (generally originating from

surveys with a small number of modalities (up to 20). The coding of curves into V.Q.O. required, in order to apply it, the development of software making it possible to calculate the reduced centered index between V.Q.O.s having a very large number of modalities (1000 and more)

2 - CODING AND COMPARISON INDEX

2.1 "Hypothèse d'absence de lien (H.A.L.)"

Given two V.Q.O.s f and g defined on the same set T and having, respectively, as modality sets:

$L = \{1,...l,...L\}$

and

$M = \{1,...m,...M\}$

on T, f and g are respectively designed by w_f et w_g respectively and are represented in T x T par $R(w_f)$ et $R(W_g)$

$$R(w_f) = \{(t,t')\epsilon T \times T / f(t) < f(t')\}$$

$$R(w_g) = \{(t,t')\epsilon T \times T / g(t) < g(t')\}$$

The raw index of similarity between f and g is defined by:

s(f,g) = Card (R $(w_f) \cap R(w_g)$)

The "hypothèse d'absence de liens (H.A.L.)" assumption is formulated as follows: the variable g is fixed and a randon variable f' is associated with the variable f in other words w' is associated with w_f in the set:

$\{n(1),...n(l),...n(L)\}$

this set is provided with a uniformly distributed probability with:

$$n(l) = Card\{t\epsilon T / f(t) = l\}$$

w' is represented in T x T by the random part R(w'), the index associated with s(f,g) is defined by:

S(g) = Card (R (w') $\cap$ R (w_g))

the mean and the variance of which are respectively:

$m_{fg} = \lambda\mu$

$$s_{fg}^2 = \lambda\mu + \rho_{11}\delta_{11} + \rho_{22}\delta_{22} + \rho_{12}\delta_{12} + (\theta\xi - \lambda^2\mu^2)$$

The expressions for:

$\mu, \delta_{11}, \delta_{22}, \delta_{12}, xi$

are respectively of the same form as

$\lambda, \rho_{11}, \rho_{22}, rho_{12}, \theta$ the first ones being relative to w_f associated with f and the second ones being relative to w_g associated with g.

If n > 4, we have:

$\lambda = \sum n(l)n(l')/1 \le l \le L/\sqrt{n(n-1)(n-2)}$

$\rho_{11} = \sum\{n(l)n1(l)(n1(l)-1)/2 \le l \le L\}/\sqrt{n(n-1)(n-2)}$

$\rho_{22} = \sum\{n(l)n_2(l)(n_2(l)-1)1 \le l \le L-1\}/\sqrt{n(n-1)(n-2)}$

$\rho_{12} = \sum\{n(l)n_1(l)(n_2(l)/2 \le l \le L-1\}/\sqrt{n(n-1)(n-2)}$

$\theta = \sum\{n(l_1)n(l'_1)\{\sum n(l_2)n(l'_2)/1 \le l_2 \le l'_2 < L\} + n(l_1) + n(l'_1) - (2n+1)/1 \le l_1 \le l_{1'} < L\}/\sqrt{n(n-1)(n-2)(n-3)}$

where $n_1(l) = \sum\{n(l')/l' \le l\}$

and $n_2(l) = \sum\{n(l')/l' \ge l\}$

The association index between f and g is put in the following reduced centered form:

$$q(f, g) = (s(f,g) - m_{fg}) / s_{fg}$$

and by reference to a scale of probability, using the distribution function of the normal law N (0,1) the final association index between f and g is written as:

$$p(f, g) = q_s(f, g)$$

where $q_s(f, g) = (q(f, g) - moy(q))/\sqrt{var}(q)$

2.2 *Coding a curve in V.Q.O.*

Let f be a curve defined on set T having values in an interval (A,B) bounded by R . We will call a quantified curve, signified by f_q , the curve defined by T having values in:

$\{L = 1,...l,...L\}$

such that, for every element t of T , we have:

$f_q(t) = l$ if and only if

$A + (l-1)(B-A)/L < f(t) < A + l(B-A)/L$

where $(B-A)/l$ is the quantification step. It is a whole number multiple of the measurement error of the equipment providing the values of f. f is said to be coded in V.Q.O. if it induces a total preorder on T .

3 - EFFECTIVE COMPARISON TECHNIQUE

This technique is relational and, contrary to the case of comparing two qualitative variables with a small number of modalities, it wasn't pertinent to have a pair the two arguments of which were too far away. Additionally, we initially limited ourselves to a local comparison of two curves, by simply comparing the intervals of strict monotony. The proposed proximity index between two curves takes the form of a sequence of indices standardized by H.A.L relating to the intervals of monotony of the first curve with respect to the second and of the second one with respect to the first. Let f and g be two curves defined on T having values in intervals bounded by R. Let ($T^1,T^2,...T^j$) (resp.($T'^1,T'^2,...,T'^k$)) be a series of intervals of T over which f (resp. g) is strictly monotonic. We shall consider $f|T^j$ and $g|T^j$ (respectively $g|T'^k$ and $f|T'^k$) as the curves coded in V.Q.O. from restrictions of f and g to T^j for j = 1,...,j (respectively to T'^k for k = 1,...k) having values respectively in:

L = {1,...,L}

and

M = {1,...,M}

The similarity index s (f,g) between the two curves f and g coded in V.Q.O. will be equal to the sum of the two following expressions:

$\sum\{card(T^j) \times q(f|T^j, g|T^j)|j = 1,...J\}/\sum\{card(T^j|j = 1,...,J\}$

$\sum\{card(T'^k) \times q(f|T'^k, f|T'^k)|k = 1,...K\}/\sum\{card(T^j|k = 1,...,K\}$

where q designates the index of association between two reduced centered V.Q.O.s with respect to the "hypothèse d'absence de lien" assumption of absence of a linkage.

4 - COMPARING THE RESULTS FROM DIFFERENT METHODS

The coding and the index we are presenting only bring the evolutional aspect of temporal data to light, independently of their amplitude. Can these two aspects be dissociated? To answer this question,

we decided to do a comparison from a purely applied point of view of the partitions of a family of curves obtained, firstly, by their classification using V.Q.O coding and, secondly, by their classification using A.V.L. taken as quantitative variables (we used correlation as the proximity index) and their classification, as individuals, using Ward's method (the index employed was the conventional euclidian distance between points in a metric space). The chosen data set consisted of 60 curves with discreet values taken in constant steps at 41 points. The curves represented a speech signal in the frequency domain. Each one of them was the smoothed spectrum associated with a phoneme's window (or instant). There were six phonemes: I AU A Z S AI.

The interest of such data is its ease of classification. An "ideal" method of classification should yield a partition of the family used into six classes each corresponding to a phoneme.

4.1 Results of classification

4.1.1 Using A.V.L. for curves considered as numeric variables

Here, (see figs. 1 and 2) a curve is considered as a numeric variable. If we use t_i to designate a point on the x-axis, t_i plays the part of the n^{th} individual and $f(t_i)$ is the variable's value on this individual. The reduced centered proximity index between two curves is proportional to the correlation coefficient.
The partition obtained is close to the "natural" partition with this method. This is confirmed by the relative "continuity" of the surface described by the two curves in each class, as can be clearly seen in the three-dimensional representation (figs. 1 and 2).

4.1.2 Using Ward's method for curves considered as individuals

Here (see figs 3 and 4) the distance used between two curves is the standard euclidian distance. The more the curves have close (or equal) values over the internal T, the more they will resemble each other. Classifying them in this way brings their resemblance at amplitude level to light.

4.1.3 Using A.V.L. for curves coded as V.Q.O. variables

This classification (see figs. 5 and 6) brings the sense of variation of the curves to light, independently of their amplitude. It will be seen that this is true if figures 5 and 6 are examined (principally figure 6), and "discontinuities" on the curves of one class will be noticed.

5 - REFERENCES

DEVILLE J.C.

Méthodes statistiques et numériques de l'analyse harmonique.

Annales de l'INSEE numéro 15 janvier - avril 74

FARAJ A.

Traitement statistique et typologie d'impulsions de foudre recueillies lors de la campagne COPELIA

Thèse de doctorat de l'université de Rennes 1 1988

FARAJ A. VALETTE N.

Analyse factorielle appliquée aux données COPELIA

Note technique NT/LAA/ITP/95 Lannion 1987

LERMAN I.C.

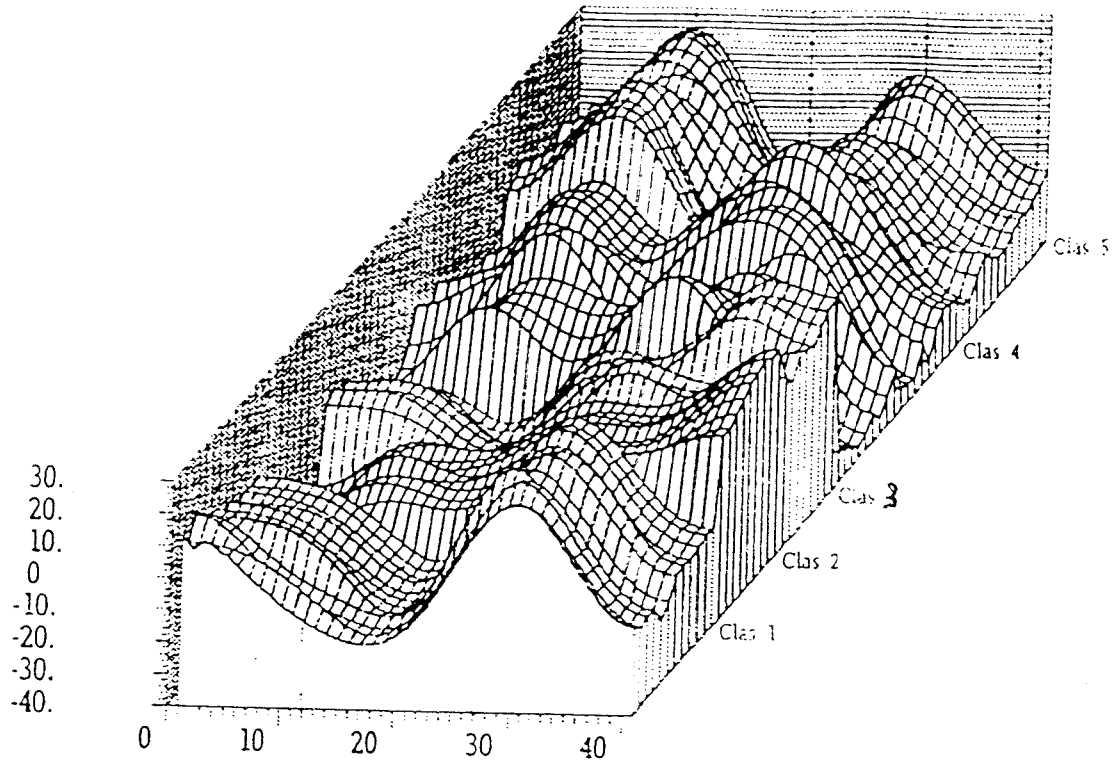
Classification et analyse ordinale des données

Dunod PARIS 1981

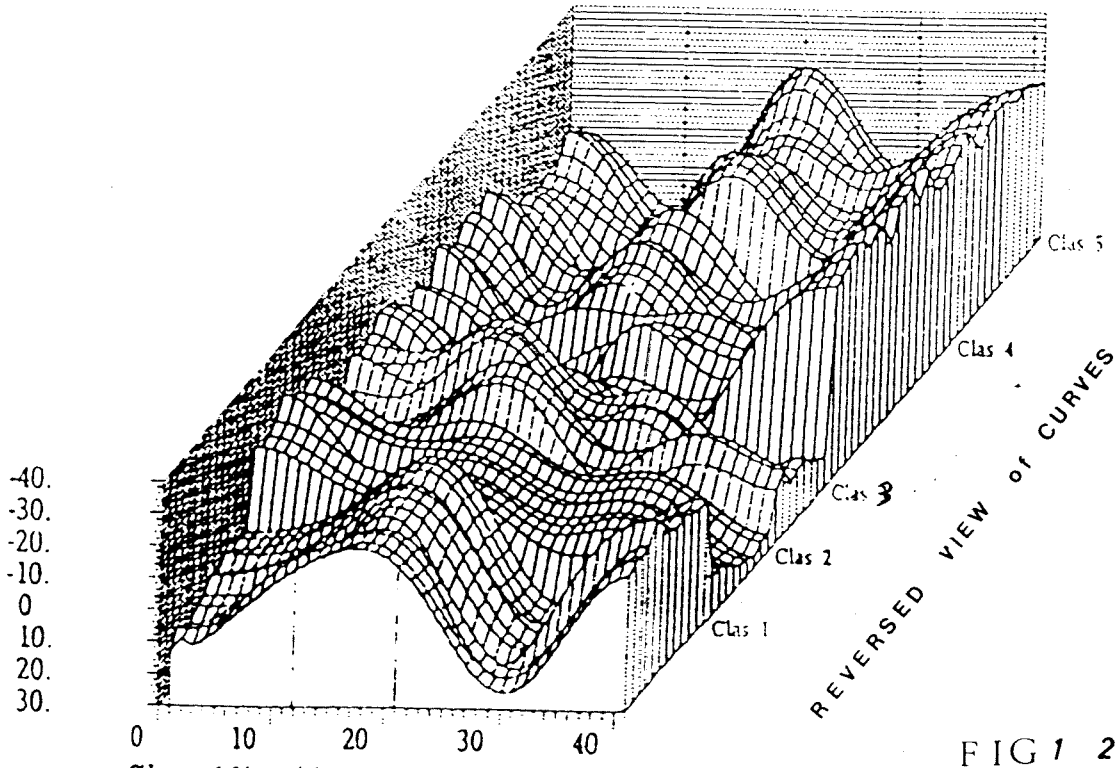
LERMAN I.C. -- PETER P.

Elaboration d'un indice de similarité entre objets de types quelconques

Rapport INRIA 1985

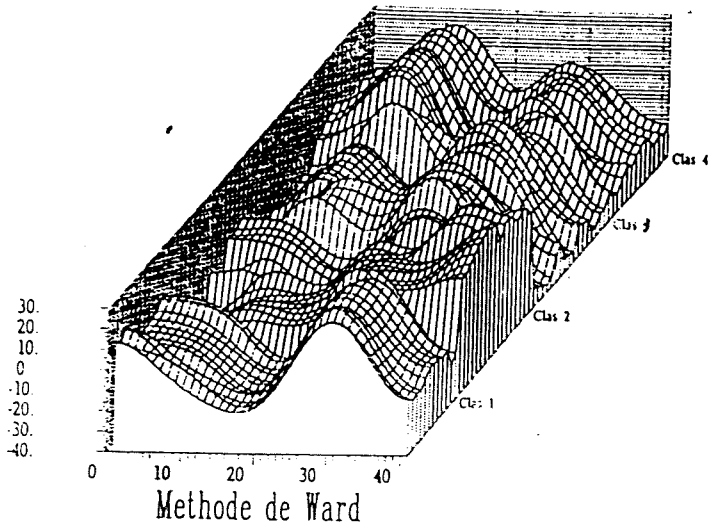


Classification par l'AVL



Classification par l'AVL

FIG 1 2



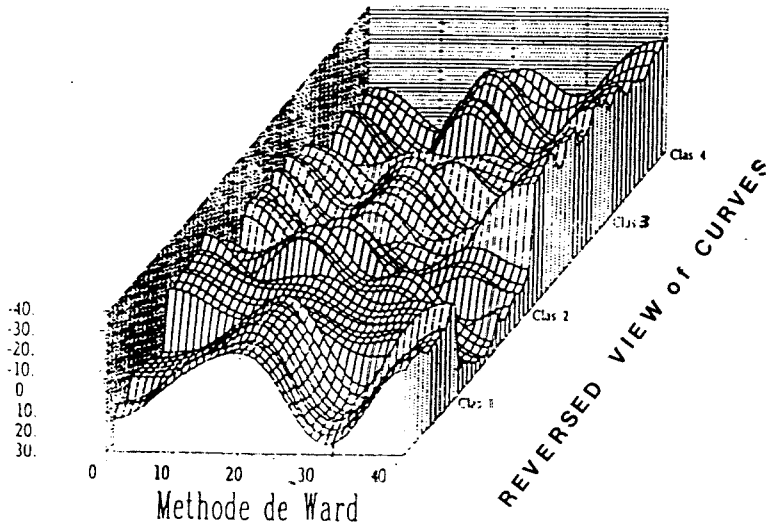
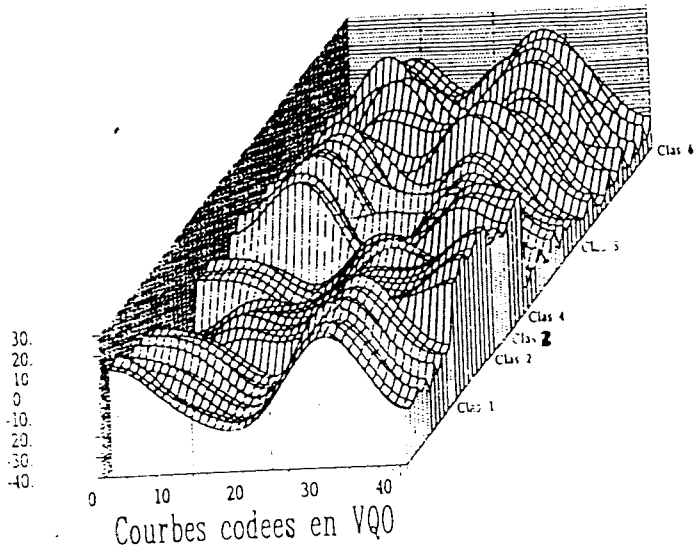
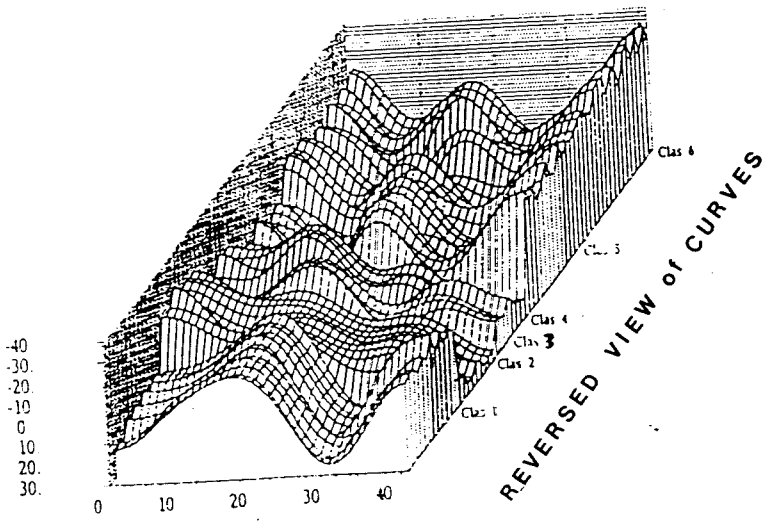


FIG. 3_4





COURBES CODEES en V.Q.O

FIG. 5_6